DOCUMENT RESUME

ED 219 402 TM 820 412

AUTHOR Frary, Robert B.
TITLE A Comparison Among Person-Fit Measures.
PUB DATE Mar 82
NOTE 28p.; Paper presented at the Annual Meeting of the American Educational Research Association and the National Council on Measurement in Education (New York, NY, March 19-23, 1982).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Higher Education; *Multiple Choice Tests; Predictive Validity; *Reliability; *Statistical Analysis; Statistical Studies; *Testing Problems
IDENTIFIERS *Person Fit Measures

ABSTRACT

Person-fit is the extent to which an examinee's response pattern on a multiple choice test is consistent with his ability as estimated by total score. Three measures of person-fit were computed for students taking classroom tests under twelve different instructors at a comprehensive university. Supplementary questions on each test inquired concerning students' health and extent of preparation at the time of testing. It had been conjectured these might be related to person-fit. Internal consistency reliability estimates were computed for the person-fit measures. Stability of these measures over time was evaluated for two sequences of tests over two groups of students. The evidence was insufficient to permit evaluating the measures with respect to each other. Other analyses suggested poor person-fit arose from an examinee's unexpected responses to a small proportion of items. Substantially different subsets of items account for this phenomenon from one examinee to another. Internal consistency reliability estimates were inappropriate as person-fit measures. These measures were determined to be unstable within examinees across different tests in the same academic course. The potential usefulness of person-fit measures requires further study. (Author/DWH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED219402

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

A COMPARISON AMONG PERSON-FIT MEASURES

Robert B. Frary

Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
R. B. Frary
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 820 412

Paper presented at the annual meetings of the American Educational Research Association and the National Council on Measurement in Education, New York, 1982.

A COMPARISON AMONG PERSON-FIT MEASURES

Robert B. Frary

ABSTRACT

Three measures of person-fit (the extent to which an examinee's response pattern on a multiple-choice test is consistent with his ability as estimated by total score) were computed for students taking classroom tests under 12 different instructors at a comprehensive university. Supplementary questions on each test inquired concerning students' physical and mental states at the time of testing (e.g., health and extent of preparation), since it had been conjectured that these might be related to person-fit. Internal consistency reliability estimates were computed for the person-fit measures, and stability of these measures over time was evaluated for two sequences of tests over two groups of students. Only very weak evidence was found in support of the conjecture that poor person-fit might be associated with examinee circumstances at the time of testing, evidence far too weak to permit evaluating the measures with respect to each other. Other analyses suggested that poor person-fit typically arises from an examinee's unexpected responses to a very small proportion of items and that substantially different subsets of items account for this phenomenon from one examinee to another. Accordingly, internal consistency reliability estimates are probably inappropriate for person-fit measures. It was also found that person-fit measures were highly unstable within examinees across several different tests in the same academic course. These results strongly question the potential usefulness of person-fit measures and call for studies to evaluate test-retest reliability of person-fit measures.

A COMPARISON AMONG PERSON-FIT MEASURES

Robert B. Frary[1]

At the level of popular discourse, hardly any condemnation of a test is more devastating than that arising in connection with a low-scoring examinee who succeeded in a venture for which the test predicted failure. Such a disclosure is inevitably followed by one of an examinee who failed miserably when the test predicted success. Measurement practitioners can accept a few such cases with equanimity but naturally seek means to minimize their occurrence. Identification of aberrant response patterns (or poor person-fit) is one approach. It is a plausible conjecture that a relatively low-scoring examinee who missed the easy questions and got the hard questions right is worthy of further consideration rather than immediate consignment to the group for whom failure is predicted.

Another testing concern is potential unfairness with respect to a subgroup of examinees. This concern has been investigated until now mainly at the item level. When a subgroup of examinees answers an item correctly significantly less or more often than their total test scores would predict, poor item-fit is said to have been observed. This approach assumes that only a small subset of items has this characteristic, since presence of a large proportion would have substantial effect on the total test scores used to establish item-fit. In contrast, a tendency toward poor person-fit within a subgroup of examinees could involve no particular subset of items and yet arise from unfairness. For example, a subgroup of examinees might tend to be highly uneasy in a testing situation and as a result make many more careless errors than other examinees. This would tend to yield worse person-fit

[1]The author wishes to extend sincere thanks to the faculty members of Virginia Polytechnic Institute and State University who collected data for this study in their classes.

within the anxious subgroup as well as scores lower than in the absence of anxiety. However, extremely poor person-fit would probably not be observed for a large proportion of the anxious subgroup; only comparison of their average person-fit with that of the residual group would reveal the problem.

For all the plausibility of the above arguments, very little has been done to validate person-fit measures or otherwise investigate their psychometric properties. To aid in reduction of this deficiency, this paper compares three substantially different person-fit measures from a variety of psychometric standpoints. In particular, validation of the person-fit measures is undertaken both with respect to examinees with extremely poor person-fit values and with respect to variation of person-fit values over examinee subgroups.

Person-Fit Measures Used

As proposed by Wright et al. (1976), for use in conjunction with the Rasch model the following statistic will tend to yield a high value for an examinee whose right-wrong response pattern is not consistent with total score and difficulty of the items:

$$\sum_i [(x_i - p_i)^2/(p_i(1 - p_i))]. \qquad [1]$$

Summation is over items for an individual examinee, and $x_i = 1$ if the examinee answers item i correctly and zero otherwise. p_i is the probability of correctness for the examinee on item i, and its estimate would be based in part on the performance of the total population on item i. The estimate may be made in various ways including use of the Rasch model. Formula [1] is approximately χ^2 distributed with N - 1 degrees of freedom. Because the denominators of the summands of [1] are small for items at the extreme levels of difficulty, such

items contribute proportionately more to the sum than under an alternate formula adopted for use with the Rasch model in the BICAL program (Wright, Mead and Bell, 1979). In this alternate formula, the numerators and denominators are summed separately prior to forming a quotient.

A study by Ryan et al. (1980), reported relatively higher values of this alternate for [1] (worse person-fit) for black and male examinees as compared with white and female. This outcome was not surprising for the black examinees, since they tended to have lower test scores, and [1] or its alternate usually correlates negatively with score. (There are relatively few opportunities for high-scoring examinees to miss easy items.) Nevertheless, a study by Frary and Giles (1980) showed the same pattern of race and sex differences in person-fit as measured by [1], even when test score level was controlled. This study also sought evidence of possible causes of the observed person-fit differences, such as more random guessing by one group than another or instances of poor item-fit for blacks or males which coincidentally contributed to poor person-fit. The results in this regard were completely negative, and the causes of the observed person-fit differences remained unexplained.

A potential weakness of [1] and its alternate is the fact that for multiple-choice tests they ignore information as to which wrong choice was selected. For example, it might be quite probable for an examinee to miss an item but quite improbable to select a specific wrong response. In Frary and Giles (1980), two statistics were derived which alleviate this weakness. These statistics are analagous to [1] and its alternate in that one sums a quotient for each item and the other sums numerators and denominators prior to forming a quotient. They were found to be somewhat intercorrelated with each other and with [1] and its alternate. However, the smallest correlations were between [1] and

the new statistic based on formation of a quotient after summation. This statistic was selected for use in the present study and is computed as follows:

$$\frac{\sum_i (p_{i_{max}} - p_{i_*}) - \sum_i [\sum_j p_{ij} (p_{i_{max}} - p_{ij})]}{\left(\sum_i [\sum_j p_{ij} (\sum_j p_{ij} (p_{i_{max}} - p_{ij}) - (p_{i_{max}} - p_{ij}))^2] \right)^{.5}} \qquad [2]$$

in which i is a subscript over items, j is a subscript over choices, p_{ij} is the estimated probability that the (unique) examinee will select choice j of item i, $p_{i_{max}}$ is the maximum probability estimate for any choice of item i and p_{i_*} is the probability estimate for the choice actually made. Effectively this statistic reflects the standardized sum across items of the differences between the estimated probabilities of the most likely choices for each item (for the unique examinee) and the ones actually marked. Given a reasonable number of items, [2] should be approximately normally distributed with a mean of 0 and a standard deviation of 1.

In the Frary and Giles (1980) study, the relationships of statistic [2] to external criteria were very similar to those for [1] in spite of the potential sensitivity of [2] to the selection of wrong choices associated with low probabilities. Thus the present study sought further evidence concerning this issue.

For both [1] and [2] probabilities were estimated by a method documented in Frary et al. (1977). For an examinee with an average score, this method assigns the observed proportion of examinees selecting each choice as the p_{ij} value. For examinees with above and below average scores, these probabilities are increased or decreased in proportion to the examinee's score. For an above-average examinee, it is assumed that the probability of using a wrong response is equal to the proportion of examinees

using that choice multiplied by the ratio of the number of wrong answers of the examinee to the average number of wrong answers. The differences between all the wrong response probabilities thus computed and the proportions selecting those responses are then added to the proportion selecting the correct answer. This and a parallel procedure for below average examinees guaranties that for any examinee the probability assigned for any response is in the range 0 to 1, that the sum of the probabilities assigned for all responses to a given item is 1 and that the sum of the probabilities assigned to the correct responses equals the examinee's score. This method was selected for its computational simplicity in the belief that estimates thus obtained could not be substantially different from those arising from other more complicated though potentially more accurate methods (mainly because of the restraints on sums of probabilities just mentioned).

The third statistic chosen for the present study is the modified version of the caution index of Sato as reported by Harnisch and Linn (1981):

$$\frac{\sum_{1}^{n}(1 - x_i)C_i - \sum_{n+1}^{N} x_i C_i}{\sum_{1}^{n} C_i - \sum_{N+1-n}^{N} C_i} . \qquad [3]$$

Summation is over items ordered according to proportion correct, n is the examinee's score, and N is the number of items on the test. C_i is the number of examinees correct on item i; $x_i = 1$ if the examinee is correct on item i and 0 otherwise. Harnisch and Linn recommended this index over others of a similar nature (not based on item probability estimates) largely in view of its relatively weak relationship with test score. In addition, they found that values from [3] based on responses to a statewide testing

program varied significantly by school, possibly as a result of inadequate subject matter coverage in schools with high average values of [3].

Data Collection

Classroom multiple-choice tests by 12 different instructors administered in the fall of 1981 at Virginia Polytechnic Institute and State University were the source of data for this study. Preliminary analyses involved one test from each instructor, to which five questions were appended concerning personal circumstances of the examinee with respect to the test. Table 1A lists these questions. They were formulated mainly to evoke responses possibly reflecting circumstances under which the test might not be appropriate for a specific examinee. The gender question was included because of past evidence that males tended to have worse person-fit. Cooperation of the examinees was requested for the purpose of aiding measurement research, and they were assured that their responses would have no effect on their grades.

Though not selected at random, the tests can be considered representative of those processed by a test scoring service in a large, comprehensive university. Some were clearly of better quality than others, but none should be considered unacceptable for its purpose. In some cases, instructors deleted items after inspecting a preliminary item analysis or receiving negative feedback from students. These items were not used in generating statistics for the study.

For reasons which are presented later, additional data were collected in the largest of the 12 classes described above (N = 500). Two open-ended questions were circulated to all students present at the second class

session following the test to which the five extra questions had been appended. These questions are shown in Table 1B. Students were told that their answers would be helpful in further analysis of their responses to the preceding test and were requested to indicate their ID numbers to make this possible. They were assured that their identities would be kept confidential and that their answers would not be shown to the instructor in handwritten form.

To investigate the stability over time of the person-fit measures, successive tests of two instructors were analyzed. One administered five tests and the other six. Each of these sets included one test to which were appended the personal questions of Table 1A.

Results and Discussion

Computation of Indices. In the computation of index [2], omission of an item must be considered a "choice" over which summation occurs. However, all the tests of this study were unspeeded and were scored on the basis of the number of correct responses, which resulted in very few omissions. Accordingly, estimates of the probability of omitting items would have been exceedingly small, which would have yielded a relatively large value of index [2] for any examinee who omitted an item (possibly due only to marking too lightly on a machine scored answer sheet). To avoid possibly spurious values of index [2], examinees who omitted any items were dropped from the study. (Indices [1] and [3] were not affected by this problem but were not computed for omitters in order to have identical groups for comparisons among the three person-fit indices.) Fewer than 1% of the available examinees were dropped.

Basic Characteristics of Indices. Table 2 presents basic test response and person-fit statistics from the 12 tests with the appended questions. Two measures of relationship among the person-fit statistics are shown, their intercorrelations and the proportion of overlap in the top 10% of cases for each person-fit statistic. These two approaches to reporting interrelationships reflect the previously stated goal of studying the person-fit measures both from the standpoint of extreme values and across all examinees. The results of Table 2 may be summarized as follows:

1) The means of index [1] are all below expectation, ranging from about 80% to 95% of their expected values. The standard deviations of index [1] seem somewhat more unstable. For example, Tests 4 and 7 with 50 and 52 items should yield approximately equal index [1] standard deviations, yet the observed standard deviations are 6.6 and 22.4 respectively, almost extremes over the 12 tests. Tests 4 and 7 had no very peculiar characteristics. Though Test 7 had a slightly low KR20, this did not reflect the presence of any items correlating negatively with total score.
2) Index [2] means are also all below expectation. These deficiencies range from about 20% to 70% of their respective standard deviations. The index [2] standard deviations are also all below expectation and seem strongly related to the mean percentage of correct answers on the tests ($r = .81$).
3) No expectations have been established for the means and standard deviations of index [3]; the smallest mean is about two-thirds of the largest, and the smallest standard deviation is slightly less

than half the size of the largest. Neither means nor standard deviations varied according to mean percentage correct on the tests to any meaningful extent.

4) All three indices tended to have rather weak correlations with test scores. Those for index [3] would certainly have to be considered negligible, with those of index [2] almost as slight. The correlations of index [1] with test scores were almost consistently negative but not of substantial size.

5) Indices [1] and [3] were substantially correlated; the mean coefficient across the 12 tests was about .75. Indices [1] and [2] were substantially less strongly related, but [2] and [3] were almost as strongly related as [1] and [3]. A similar pattern occurred for percentage of overlap in the top 10% on each index.

Relationship of Indices to Criteria. Responses to appended questions 1, 2, 4 and 5 (Table 1A) were scaled ordinally according to their choice numbers. Question 3 was scaled 1 = fair 2 = unfair, since only rather small proportions of examinees selected any particular reason for unfairness. (These choices were included mainly to aid the cooperating instructors in interpretation of negative responses.) An additional variable was created by subtracting the value for question 2 from that of question 1; this represented the degree of departure from the usual amount of studying for the test in question with positive values indicating more study than usual and negative values less. Only a very small proportion of examinees in any class omitted any of the appended questions.

For each of the 12 examinations of Table 2, each of the six personal

response variables just described was correlated with each of the three indices, thus generating 18 correlation coefficients per examination. In addition, personal responses of the top 10% on each index were compared with those of the lower 90% using χ^2 tests, which generated 18 χ^2's for each examination. Resulting correlation coefficients and χ^2's were inspected for their statistical significance with the intention of interpreting probability values extremely conservatively given the large number of statistical tests.

Table 3 reviews the results of the 18 χ^2 tests and 18 correlation coefficients for Test 3 from Table 2. As shown there, eight of the 36 statistical tests yielded probabilities of less than .05, and five of these involved the fairness variable (1 = fair, 2 = unfair). While weak, the significant correlation coefficients do suggest a tendency for those perceiving the test as unfair to have higher (worse) person-fit scores on all three indices. Cell proportions for the significant χ^2 tests related to the fairness variable reflect the same tendency: proportionately more subjects in the top 10% of person-fit viewed the test as unfair than in the bottom 90%. Cell proportions for the other three significant χ^2 tests suggested that those in the top 10% had greater tendencies not to have studied much (question 1), to have been affected more by personal problems (question 4) and to have studied less than usual (question 1 minus question 2).

Unfortunately, the results of Table 3, while moderately supportive of conjectures concerning person-fit, are not typical of the tests of this study. Over the 396 statistical tests associated with the remaining

11 tests, only 23 yielded probabilities of less than .05. Of these 11 tests, 10 had three or fewer probabilities of less than .05, and one had five. Thus, given the previously stated commitment to conservative interpretation of significance levels, it is not possible to claim any very strong or pervasive relationship between person-fit and self-reported examinee circumstances. (Of course, this denial refers only to the specific topics questioned, the types of tests used and the population sampled.) Nevertheless, it is interesting to note that all but two of the 31 significant statistical tests suggested the sort of relationships conjectured earlier with respect to person-fit, e.g., high (poor) person-fit with lack of studying, personal problems, perceived unfairness or being male. The two exceptions were with respect to gender. Females tended to have worse person-fit on Test 7 as evidenced by a correlation of .23 between sex (m = 1, f = 2) and both index [2] and index [3].

Analysis of Openended Responses. The largely negative results just described motivated the collection of openended responses. (These were from the class which had taken Test 4.) It seemed possible that the lack of relationships might be due to the restricted nature of the appended questions and that less focused stimuli might evoke responses revealing possible reasons for poor person-fit.

Of the approximately 500 students in this class, 328 responded to the questions of Table 1B and gave ID numbers matching ones from the preceding test. The researcher and two colleagues each read responses from 50 students and agreed on the existence of six response categories or themes to which nearly all of the comments could readily be assigned

irrespective of which of the two questions gave rise to them. These response categories are listed in Table 4 along with the percentage of responders assigned to each by the researcher. Only a very small number of responses could not be categorized. Many students made multiple comments; hence the percentages sum to more than 100. Only about 10% answered both openended questions negatively.

While some of the response categories of Table 4 are essentially the same as those of the appended questions, the response levels are quite different. For example, only 9% of the responders to the openended questions mentioned the effect of personal problems, such as illness, while 76% claimed at least some effect from personal problems on appended question 4. Other openended response categories were quite unrelated to those of the appended questions.

The six response categories of Table 4 were coded as dichotomous variables and analyzed in the same manner as the variables from the appended questions, thus yielding 18 correlation coefficients (6 variables x 3 indices) and 18 χ^2's. Of the 36 associated statistical tests only three had probabilities of less than .05, thus casting substantial doubt on the conjecture that narrowness of content in the appended questions was responsible for the lack of relationship between responses to them and person-fit.

Analysis of Successive Tests. While some of the circumstances which might evoke poor person-fit could be transitory (e.g., questions 1, 3, 4 of Table 1A), others would not (e.g., questions 2, 5), and even those which are transitory could prevail with certain examinees. Thus, persistence of

poor person-fit across several tests would be of substantial interest. To determine the extent of this phenomenon, the two sets of successive tests described earlier were analyzed to produce correlations between person-fit values for a given index for each pair of tests within a set. Thus for the set of five tests (which included Test 2 of Table 2), three 5 x 5 correlation matrices were produced, one for each index; three 6 x 6 matrices were produced for the other set (which included Test 3 of Table 2). Table 5 summarizes these results by reporting the maximum, minimum and average correlations between administrations within each index. In addition, lists of examinees in the top 10% for each index on each test were inspected to determine the number of times specific examinees were in the top 10%. These results are also reported in Table 5.

These outcomes certainly do not speak for any substantial persistence of poor person-fit across successive testings for the tests and population studied. Indeed, over all three indices and both sets of tests, there were only ten intertest correlations (out of 75) significantly different from zero ($p < .05$). Nevertheless the finding of an examinee in the top 10% of index [2] values on four tests from the five test set was of interest, since the probability of this event (in isolation) is only about .03. The instructor was asked whether this student displayed any unusual characteristics. The answer was negative.

Internal Consistency Reliability Analysis. The results described thus far led to the question of whether their suggestive but far from conclusive nature might arise from low reliability. (Apparently the question of reliability of person-fit measures has not been reported in the literature.)

Accordingly, an odd-even, split-half reliability estimate was computed for each index on each of the 12 tests of Table 2. (Coefficient α could be computed only for index [1] and hence is not reported.) Results are shown in Table 6. The extremely low (even negative) values encountered for some tests required further investigation. A first step was simply to display for indices [1] and [2] the item values summed to obtain the person-fit scores for each examinee. (These values were unstandardized for index [2] but nevertheless revealed the items contributing more heavily to high person-fit scores.) Inspection of these subject-by-item matrices revealed that a large proportion of high person-fit values arose from the examinees' unexpected responses to very small numbers of items. For example, a high scoring examinee might miss an item answered correctly by over 90% of the examinees, or an examinee with a moderate score might mark a wrong choice along with only one or two other examinees on an item of average difficulty. Given the way the indices are computed and any reasonable basis for estimating probability of answering correctly (index [1]) or selecting a specific choice (index [2]), only two or three really unusual responses are needed on a 25-item test to generate a very high person-fit score. Moreover, among those with high person-fit values, there was very little consistency as to which items were major contributors. Of course, this explains the low split-half coefficients; the small number of items contributing strongly to any particular high person-fit score could easily be found in either half of any split of the test items.

The prevalence of high person-fit values associated with very small numbers of highly aberrant responses gave rise to consideration of whether

these events might have represented random error. Accordingly, several of the tests with larger numbers of examinees were reanalyzed dropping items with fewer than 3% of examinees on any active choice. From about one-fourth to one-half of items were dropped on this basis. This procedure did yield higher reliability estimates in every case, but only slightly higher, even after use of the Spearman-Brown formula to adjust for decrease in test length. Thus the low internal consistency reliability of the tests analyzed cannot readily be attributed to some peculiarity of the present data. Also, this elimination of items yielded no stronger relationships between person-fit and the appended questions than before.

Conclusions

Based on the analyses just reviewed, a recommendation of extreme caution in the use of person-fit statistics seems highly justifiable. While faint evidence of the validity of these statistics did arise, it would not be unreasonable to conclude that, within the limits of generalizability for this study, the potential utility of these statistics has not been demonstrated. Were it not for the consistent behavior of the indices with respect to each other across tests, one might be tempted to characterize the outcomes as little different from chance occurrences. Nevertheless, other studies have yielded consistent and meaningfully large variations in person-fit statistics with respect to sex, race and instructional conditions (Ryan et. al., 1980; Frary and Giles, 1980; and Harnisch and Linn, 1981).

The present lack of positive results may be due in part to the nature of the population sampled. This population was relatively homogeneous

with respect to academic ability and contained only a very small proportion of nonwhite examinees. Moreover, within tests, there were no differences in instructional treatment. The studies previously mentioned involved populations differing with respect to one or more of these characteristics.

In view of the results of this study, there seems to be considerable need for further investigation of the psychometric properties of person-fit statistics. Issues that should be addressed are:

1) Test-retest reliability of person-fit measures. If, as in this study, high person-fit scores are consistently found to arise from unusual responses to small numbers of items differing substantially from examinee to examinee, this approach will be the only valid one for establishing the most fundamental criterion for usefulness of these statistics.
2) The effects of variation in population heterogeniety on variables related both to the test proper (e.g., academic treatment) and other personal characteristics such as socioeconomic status.
3) The effects of differing methods for estimating the probability of a correct answer or of selecting a specific choice. For example, it might be desirable to adopt some Bayesian procedure for adjusting extreme probability estimates arising from relatively small samples.

A basic concern of the present research was the extent to which three rather differently conceived person-fit measures performed the same task, namely, identification of examinees for whom a test was inappropriate. For the present data, it is necessary to conclude (conservatively) that none of

them performed this task or (more liberally) that all gave some evidence of performing the task but such weak evidence that distinguishing among them is not possible. Nevertheless, it was interesting to note the higher correlations between indices [1] and [3] than for either of them with index [2] in spite of the conceptual similarity between indices [1] and [2]. Indices [1] and [3] are, however, both based essentially on right versus wrong answers to each item, while [2] uses information from the distractors. At the least, this outcome suggests the need to include measures like index [2] in further studies comparing person-fit measures.

Finally, the research reported above suggests that use of person-fit measures for any decision-making purpose, especially with respect to individual examinees, should be undertaken only with extreme caution and that substantial additional research will be required before they can be used routinely.

REFERENCES

Frary, R.B., Tideman, T.N. and Watts, T.M. Indices of cheating on multiple-choice tests. *Journal of Educational Statistics,* 1977, *2,* 235-256.

Frary, R.B. and Giles, M.B. Multiple-choice test bias due to answering strategy variation. Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, Massachusetts, 1980.

Harnisch, D.L. and Linn, R.L. Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement,* 1981. *18,* 133-146.

Ryan, J.P., Garcia-Quintana, R., and Hamm, D.A. Testing the fit of subjects to a latent-trait model. Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, Massachusetts, 1980.

Wright, B.D., Mead, R.M. and Bell, S. *BICAL: Calibrating items with the Rasch model.* (Research Memorandum No. 23/3) Chicago: Statistical Laboratory, Department of Education, University of Chicago, 1979.

Wright, B.D., Mead, R.M. and Draba, R. *Detecting and correcting test item bias with a logistic response model.* (Research Memorandum No. 22) Chicago: Statistical Laboratory, Department of Education, University of Chicago, 1976.

Table 1 - Questions Concerning Personal Circumstances with Respect to Tests

A. Questions Appended to Tests in 12 Classes

1. How much did you study for this test?

   1) Not very much
   2) A moderate amount
   3) Very much

2. How much do you usually study for a test in a class of this type?

   1) Not very much
   2) A moderate amount
   3) Very much

3. How do you rate the fairness of this test? (Select one answer which best characterizes this test.)

   1) Not unfair
   2) Unfair--too many questions not related to assignments or lectures
   3) Unfair--questions too difficult
   4) Unfair--questions poorly written or confusing

4. To what extent do you believe personal problems (health, emotional, unusual obligations, etc.) harmed your performance on this test?

   1) Substantially
   2) Moderately
   3) Slightly
   4) Not at all

5. Sex:

   1) Male
   2) Female

B. Openended Questions Administered in One Class

1. Do you think that you did <u>less</u> <u>well</u> on the last test in General Biology than you could have or should have?

   Yes ___ No___ If yes, please explain.

2. Was there anything unexpected or unusual as far as you were concerned with the last test in General Biology?

   Yes ___ No___ If yes, please explain.

Table 2 - Test Score and Person-Fit Statistics

Test Number	1	2	3	4	5	6	7	8	9	10	11	12
Number of Examinees	78	79	188	498	94	86	100	73	401	77	174	141
Number of Items	36	36	66	50	50	25	52	35	40	37	25	26
KR20	.68	.70	.77	.82	.70	.70	.56	.43	.73	.70	.73	.65
Mean % Correct	80	72	73	68	75	74	77	73	65	69	68	53
Index [1]												
Mean	28.9	31.4	54.2	44.4	37.8	21.0	43.7	30.3	35.1	31.7	20.2	24.0
S.D.	9.2	11.3	18.7	6.6	11.2	9.2	22.4	11.6	8.5	5.0	5.8	4.0
Correlation with test score	-.20	-.23	-.19	-.27	-.50	-.08	.08	-.04	-.20	-.21	-.36	-.19
Index [2]												
Mean	-.16	-.20	-.33	-.15	-.21	-.16	-.35	-.25	-.26	-.18	-.35	-.15
S.D.	.34	.45	.51	.47	.45	.37	.51	.56	.62	.45	.51	.73
Correlation with test score	-.24	.15	.01	-.14	-.30	-.05	.14	-.00	-.04	.08	-.12	-.09
Index [3]												
Mean	.31	.24	.23	.29	.24	.25	.21	.24	.23	.27	.21	.26
S.D.	.13	.11	.08	.09	.09	.15	.07	.10	.09	.10	.07	.10
Correlation with test score	-.25	-.02	.05	.14	-.17	.04	.11	.02	-.14	.03	.01	-.12
Inter-correlations												
[1],[2]	.46	.42	.58	.49	.63	.49	.46	.65	.47	.51	.56	.51
[1],[3]	.77	.67	.66	.74	.73	.71	.60	.81	.80	.76	.86	.88
[2],[3]	.66	.60	.82	.58	.71	.75	.84	.77	.60	.63	.63	.58
% Overlap Top 10%												
[1],[2]	43	29	47	39	44	25	22	43	45	14	53	36
[1],[3]	71	29	65	67	33	63	56	71	58	43	53	71
[2],[3]	71	29	53	35	22	38	56	57	55	14	41	29

Table 3 - Statistics Relating Person-Fit Scores to Personal Variables from Test 3

Variable	Range	Mean	Correlation with Index [1]	Correlation with Index [2]	Correlation with Index [3]	χ^2 top 10% versus bottom 90% Index [1]	χ^2 top 10% versus bottom 90% Index [2]	χ^2 top 10% versus bottom 90% Index [3]
Question 1 Amount of Study N = 172	1 to 3	2.37	-.13	-.31	-.12	7.30*	.84	2.06
Question 2 Usual Amount Of Study N = 173	1 to 3	2.42	-.05	-.03	-.10	.73	.95	.73
Question 3 Fair versus Unfair N = 179	1 to 2	.16	.19*	.15*	.20*	2.69	9.28*	5.50*
Question 4 Personal Problems N = 173	1 to 4	2.79	-.14	-.11	-.10	7.51	9.19*	2.86
Question 5 Sex N = 171	1 to 2	1.53	-.08	.01	-.05	.00	.28	1.10
Question 1 Question 2 Study Discrepancy N = 172	-2 to 2	-.05	-.08	-.10	-.02	8.84*	2.10	1.24

* $p < .05$

Table 4 - Openended Response Categories
for the Questions of Table 2B

Category	Percentage of Responders (N = 528)
1. Mention of a personal problem which hampered test performance	9
2. Complaint that questions for which all answer choices had some element of correctness were unfair	11
3. Open admission of failure to study sufficiently	20
4. Protestation to the effect of having studied a great deal only to perform poorly	32
5. Complaints that problem-solving questions were unfair in that nothing could be memorized which would assure a correct answer	35
6. General non-specific complaints about difficulty of test	38

Table 5 - Correlations and Overlap in Top 10% Across Repeated Tests Within Each Person-Fit Index

	Index [1]	Index [2]	Index [3]
Five Test Set (N ≈ 75)			
Maximum correlation	.36	.28	.22
Minimum correlation	-.21	-.12	-.16
Average correlation[1]	.05	.11	.01
No. of examinees in top 10% on 4 tests:	0	1	0
3 tests:	0	0	0
2 tests:	2	8	6
Six Test Set (N ≈175)			
Maximum correlation	.25	.23	.21
Minimum correlation	.00	-.10	-.08
Average correlation[2]	.11	.04	.06
No. of examinees in top 10% on 4 tests:	0	2	0
3 tests:	0	4	1
2 tests:	16	6	14

[1] 10 Coefficients
[2] 15 Coefficients

Table 6 - Odd-Even Split-Half Reliability Estimates

Test No.	Index [1]	Index [2]	Index [3]
1	[1]	[1]	.15
2	[1]	[1]	.15
3	.10	.21	.23
4	.23	[1]	.21
5	.04	[1]	.01
6	.03	.25	[1]
7	.20	.05	[1]
8	.37	.53	.33
9	.28	.09	.35
10	.21	.21	.23
11	.22	.11	.20
12	[1]	.17	[1]

[1]Negative correlation between halves.